Wide Area Information Servers:
An Executive Information System for Unstructured Files

Brewster Kahle, Harry Morris, Franklin Davis
Thinking Machines Corporation

Thomas Erickson
Apple Computer

Clare Hart
Dow Jones & Co.

Robin Palmer
KPMG Peat Marwick

November 1991

Abstract

In this paper we present a corporate information system for untrained users to search gigabytes of unformatted data using quasi-natural language and relevance feedback queries. The data can reside on distributed servers anywhere on a wide area network giving the users access to personal, corporate, and published information from a single interface. Effective queries can be turned into profiles, allowing the system to automatically alert the user when new data is available.

The system was tested by twenty executive users located in 6 cities. Our primary goal in building the system was to determine if the technology and infrastructure existed to make end-user searching of unstructured information profitable. We found that effective search and user interface technologies for end-users are available, but network technologies are still a limiting cost factor.

As a result of the experiment we are continuing the development of the system. This paper will describe the overall system architecture, the implemented subset, and the lessons learned.

1. Introduction

Systems that allow corporate executives to access personal, corporate, and published information such as memos, reports, manuals, and news are new in the field of information management. The first integrated systems are just now coming on the market. They exploit networking, online mass storage, and end-user search systems, each of which has existed for some time, but their combination and integration has not been available for the corporate environment.

Commercial systems exist in each of the personal, corporate, and published data areas with different levels of user friendliness. ON Location™, for instance, allows easy content based retrieval of personal files on a Macintosh, while Lotus Megellean™ performs a similar function on a PC. Verity's Topic™ system allows for searching of LAN-based (usually corporate) archives but primarily for a trained user community. Dialog, Dow Jones, and Mead Data are major online providers of published information, but again the majority of their users are professionals in the field of information retrieval (such as corporate librarians).

Academic systems have also been developed for some of these applications. The Information Lens project [Malone 86] revolves around structured electronic mail to help in automatic organization and retrieval of business information. Project Mercury [Ginther-Webster 90] is a remote library searching system that uses a client-server model. The Smart system [Salton 71] is a information retrieval system that embodies many different searching strategies. The SuperBook project [Egan 89] is working on user interfaces for information systems with a concentration on the scientific user. Each of these systems is pushing new ground, but do not cover a complete solution for the business executive in searching diverse information sources.

The Wide Area Information Servers (WAIS, pronounced "ways") system was constructed to test the acceptability of an integrated search system directly targeted to executives [Kahle 89]. The companies participating in the project offered expertise in different parts of the problem: Dow Jones, with its business information sources; Thinking Machines, with its high-end information retrieval engines; Apple, with its user interface background; and KPMG Peat Marwick, with its information-hungry user base. Through this project, we wanted to find if the wide area information retrieval market could incorporate more users than the

trained searchers who are familiar with a variety of query languages and databases.

In the WAIS project we used a general architecture and built a small implementation to test the feasibility of an integrated information retrieval system for corporate end users. This paper is a report on the overall architecture, the various implementations, and the lessons learned from this work.

2. The WAIS Architecture

The WAIS system took advantage of available technology to make a system which could be tested on corporate executives to determine user acceptability. The system was composed of: clients, servers, and the protocol which connects them. The information servers were Connection Machine systems, running a parallel signature based search algorithm [Stanfill 86]. The cross country network connected several LANs with leased lines running the AppleTalk and TCP carrying a variation on the Z39.50 application protocol. The clients ran on Macintoshes. This section will describe the overall architecture, and the next section will describe what exactly was implemented and used during the experiment.

The WAIS architecture had the following goals:

- accessible to novice users - little or no training should be required in order to perform effective searches.
- remotely accessible - the servers must be accessible over a variety of networks.

- uniform interface - a variety of databases whether personal, corporate or published must be accessible from the same user interface.
- automatic alerting - it must be easy to create profiles for background searching
- scalable - the system must scale in number of servers, size of servers, and intelligence of servers.
- security - individuals and groups should be able to maintain control of who accesses their data.
- pricing model - a variety of information pricing structures, from per-minute charges, to subscriptions must be supported.
- multimedia - the system must support the retrieval of any file format.

Many of these goals were achieved, while others, such as pricing model experimentation, were left unresolved.

In a client-server system, the client program is the user interface, the server does the searching and retrieval of documents based on indices, and the protocol is used to transmit the queries and responses. The client and server are isolated from each other through the protocol so that they can be physically distant and interchangeable. Any client which is capable of translating a user's request into the standard protocol can be used in the system. Likewise, any server capable of answering a request encoded in the protocol can be used. In order to promote the development of both clients and servers, the protocol specification is public, as is its initial implementation.

On the client side, searches are formulated as quasi-natural language questions. The client application then formats the query for the WAIS protocol, and transmits it over a network to a server. The server receives the transmission, translates the received packet into its own query language, and searches for documents satisfying the query. The ranked list of relevant documents are then encoded in the protocol, and transmitted back to the client. At this point, the servers do not "understand" the quasi-natural language question posed by the user in any sense that a human would, but it uses the words and phrases in the question to find documents that use those terms. The client decodes the response, and displays the results. Documents of interest to the user can then be retrieved from the server.

2.1 Searching

We modeled the searching strategy on the interactive process people use when talking with a reference librarian. The library scenario is one where the client approaches a librarian or researcher with a description of needed information. The librarian might ask a few background questions, and then draw from appropriate sources to provide an initial selection of articles, reports, and references. The client sorts through this selection to find the most pertinent documents. With feedback from these trials, the researcher can refine the materials and even continue to supply the user with a flow of information as it becomes available. Monitoring which articles were useful can help the researcher provide appropriate information in the future.

The WAIS system uses a similar means of interaction: the user states a question in unrestricted natural language to a set of sources, and a set of document descriptions is retrieved. The server assigns each document a score, based on

how closely the words in the document matched the question. The user can examine any of the documents, print them, or save them for future use. If the initial response is incomplete or somehow insufficient, the user can refine the question by stating it differently.

Once a good document is found, the user may say "I want more like this one" by marking the retrieved documents as being "relevant" to the question at hand, and then re-running the search. This method of query refinement is called "relevance feedback" [Salton 83]. The server uses the marked documents to attempt to find others which are similar to them. In the present WAIS server, "similar" documents are ones which share a large number of statistically significant words and phrases. This brute force method works surprisingly well with large collections of documents [Stanfill 91] [Stanfill 86].

2.2 A Common Protocol for Information Retrieval

One of the most far reaching aspects of this project was the development of an open protocol. The four companies jointly specified a standard protocol for information retrieval by extending an existing public standard, Z39.50-1988 [NISO 88]. We choose this public standard rather than inventing one ourselves since it was close to what we needed and it could help us keep the protocol from being regarded as proprietary.

The use of an open and versatile protocol can foster hardware independence and competition. This not only provides for a much wider base of users, it allows the system to evolve over time as hardware technology progresses. For example, the protocol provides for the transmission of audio and video as well as text, even though at present most personal computers are unable to handle them. However, they are free to ignore pictures and sound returned in response to question, and to

display and retrieve only text, if that is all they are capable of processing. Higher-end platforms are free to exploit their greater processing power and network bandwidth.

Z39.50 is a general attribute based boolean search protocol intended to run over the OSI stack. It was designed for search and retrieval of bibliographic (Marc) records in libraries. As such its structure allows easy access to traditional boolean search systems such as STAIRS [Salton 83].

The WAIS protocol is an extension [Davis 90] of the existing Z39.50-1988 standard but we are working with the standards committee to merge the extensions back into the newer versions. The extensions allow support for multi-media data, large documents, a directory of servers, different communication systems, and distributed retrieval. To support multi-media, it was necessary for a document to advertise the list of available representations that could be retrieved. This was accomplished by listing the set of available types in the search response from which the client can choose one to retrieve. Another problem with the protocol involved retrieving large records. Large documents, whether text or not, would be slow to display if the whole document had be retrieved at one time, as is required in the original standard. Large documents are supported in the WAIS protocol by allowing the client to retrieve sections of a document based on bytes or lines. We also standardized a format for describing servers [Kahle 91-1] and how to contact them, which is necessary to implement a directory of servers. To support communication systems other than the full OSI protocol stack, a header was needed to show how long the packet was and how it was encoded. With this packet header we implemented the WAIS protocol over modems, TCP/IP, and X.25 systems. To support distributed retrieval we needed a document identifier system that could be used in a distributed environment [Kahle 91-2].

The protocol used in the WAIS system has proven useful in the distributed full-text environments that we tested it in.

2.3 User Interfaces: Asking Questions

Users interact with the WAIS system through the Question interface. Each question form has an area to enter the user's quasi-natural language question, the list of sources which will be accessed to try to answer the query, the list of relevant documents, and a list of answer documents.

The illustrations here are taken from the initial WAIStation program produced at Thinking Machines for the Apple Macintosh. We have also built clients for X windows and gnu-emacs. Another Macintosh interface was developed that emphasizes the alerting feature [Erickson 91].

With most current retrieval systems, complications develop when one begins dealing with more than one source of information. For example, one contacts the first source, asks it for information on some topic, contacts the next source, asks it the same questions (most likely using a different query language, a different style of interface, a different system of billing), contacts the next source, and so on. One of the primary goals behind the development of the WAIS system was to replace all this with a single interface.

With WAIS, the user selects a set of sources to query for information, and then formulates a question. When the user presses the RUN button (see figure 2), the system automatically asks all the desired servers for the required information with no further interaction necessary by the user. The documents returned are sorted and consolidated in a single place, to be manipulated by the user. The user has transparent access to a multitude of local and remote databases.

From the user's point of view, a server is a source of information. It can be located anywhere: on the local machine, on a network, or on the other side of a modem. The user's workstation keeps track of a variety of information about each server. The public information about a server includes how to contact it, a description of the contents, and the cost. In addition, individual users maintain their own private information about the servers they use.

Users may need to budget the money they are willing to spend on information from particular servers, know how often and when each server is contacted, and assess the relative usefulness of each server. In the current interface, the budget entries were put in as placeholders, since all servers are currently free. How often a source should be contacted, however, is supported but is only useful in cases where servers are updated with new information. When a source is contacted, all questions that refer to the source are updated with the new results.

A "confidence factor" allowed users to multiply the score returned from different servers so that the list presented to the user would be more appropriate. This was put in the interface to anticipate a number of different server technologies with different scoring algorithms. The "confidence factor" allows the user to adjust the scores. Also, a user might have a preference of the information from one server over another so a subjective balance would be helpful. This feature was rarely, if ever, used since the number of servers was small, they all used the same server technology, and most users only asked one source at a time anyway.

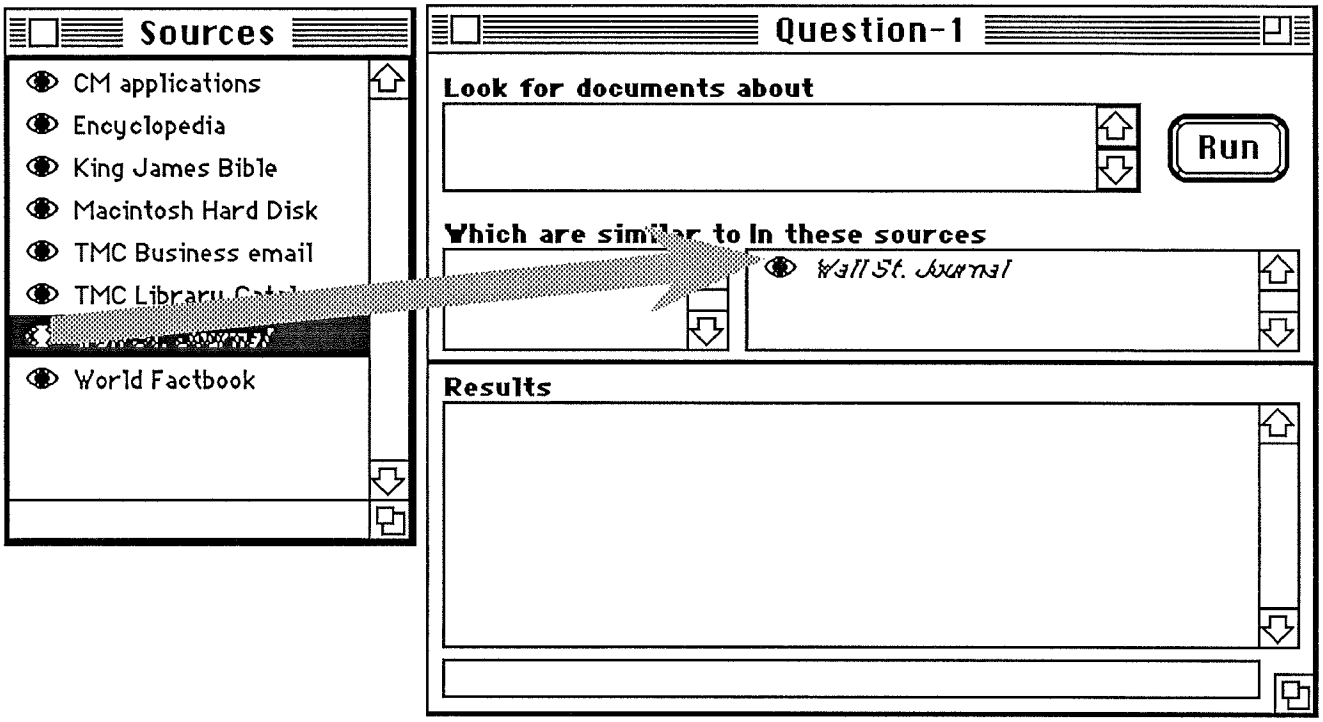

Figure 1: Sources are dragged with the mouse into the Question Window. A question can contain multiple sources. When the question is run, it asks for information from each included source.

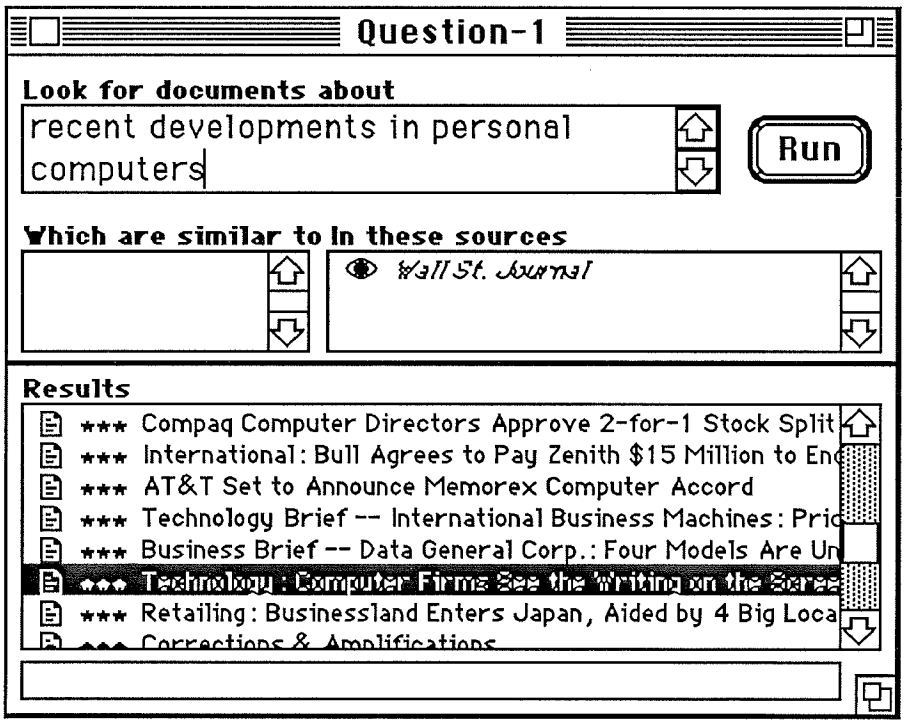

Figure 2: When a query is run, headlines of documents matching the query are displayed.

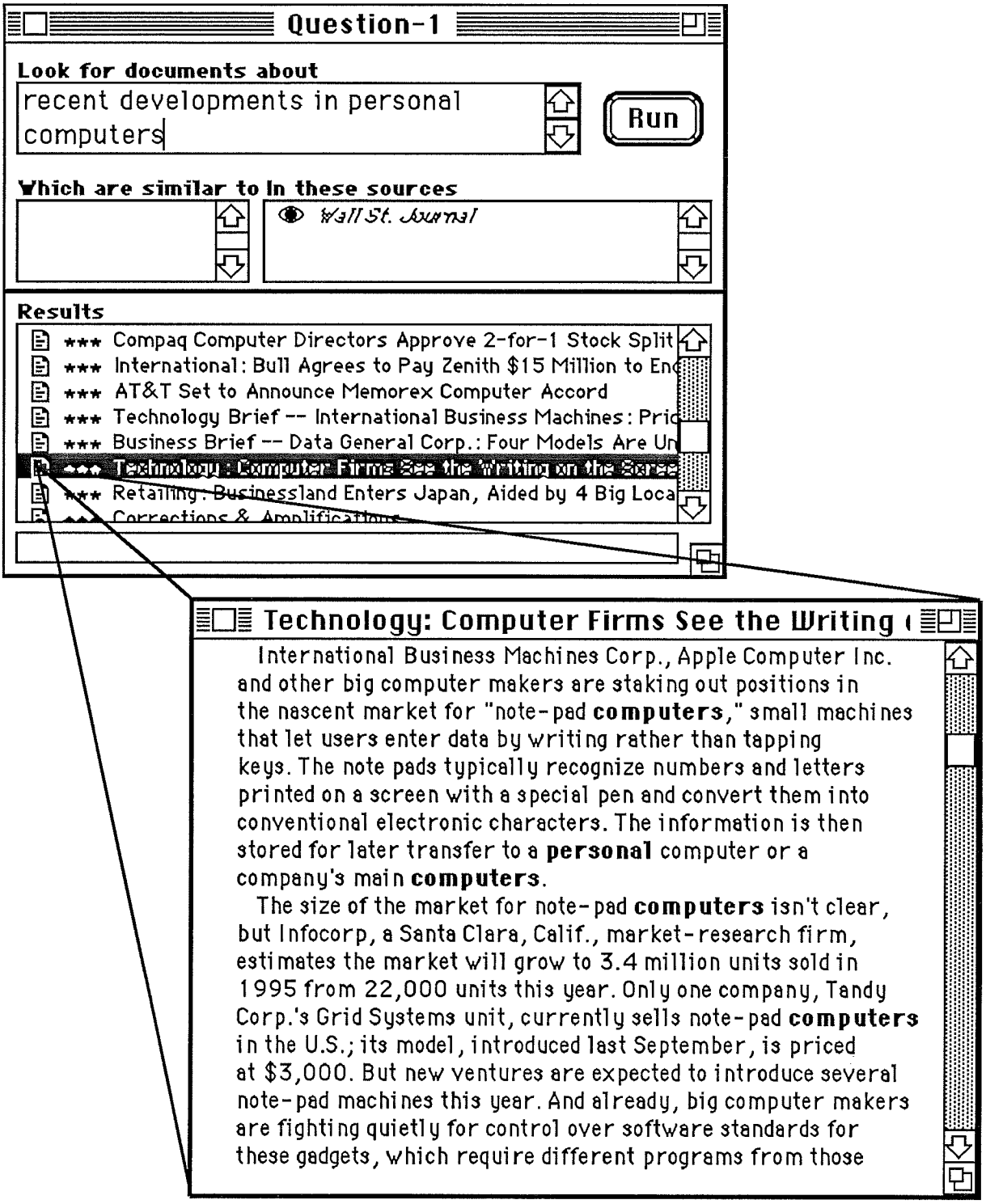

Figure 3: With the mouse, the user double clicks on any resulting document to retrieve it. The document can be text or graphics.

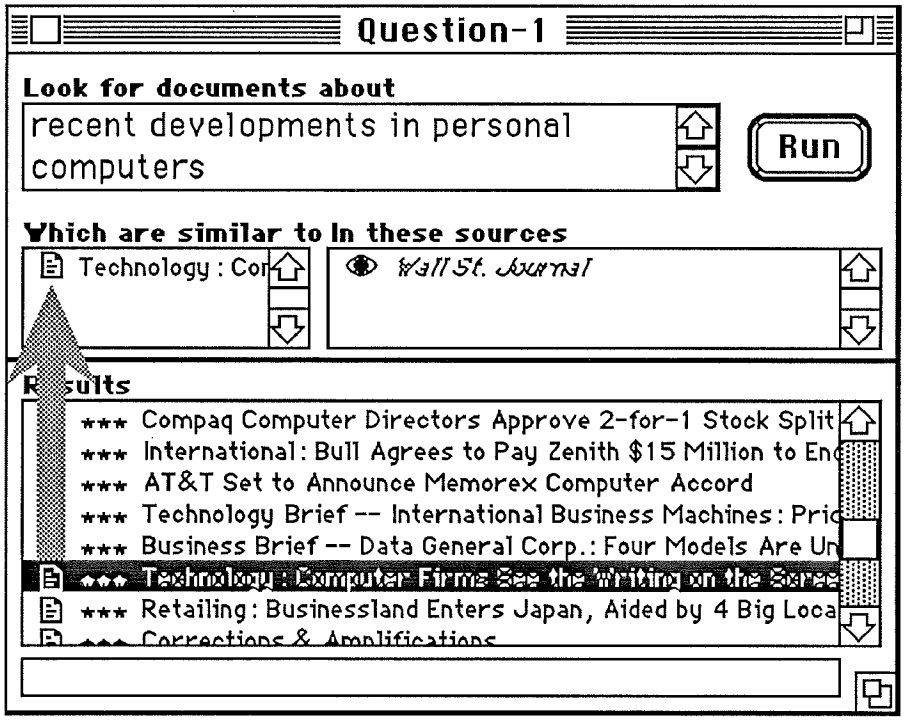

Figure 4: *To refine the search, one or more of the result documents can be moved to the "Which are similar to:" box. When the search is run again, the results will be updated to include documents which are "similar" to the ones selected.*

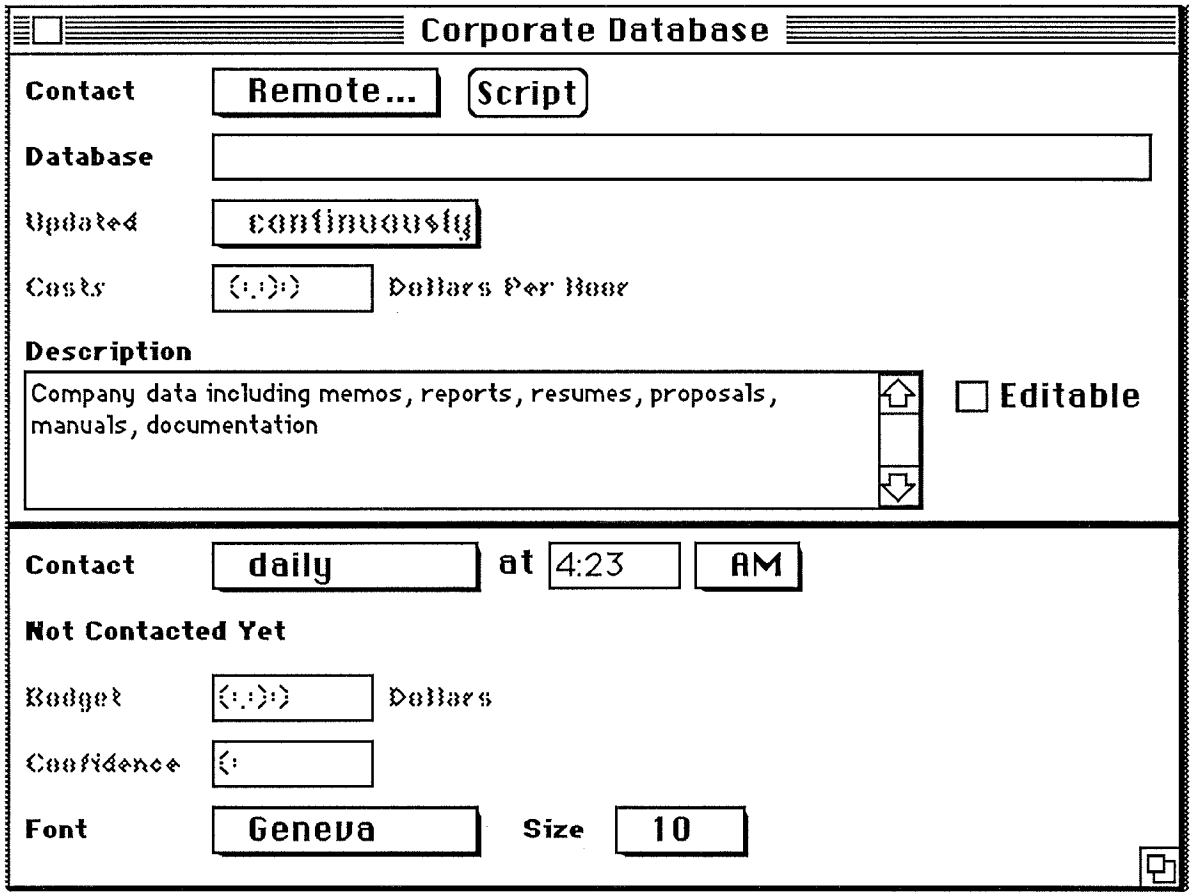

Figure 5: The Source description contains all the necessary information for contacting an information server.

2.4 Servers

The servers in the WAIS system hold a database that can be queried by a client. References of documents that best match the words and phrases in the query are returned to the client. A client can then request all or part of a document from the server. Since the client explicitly contacts the server, any number of billing methods could be employed such as 900 numbers, credit cards, and subscriptions.

The Connection Machine server system (CMDRS), used in the WAIS system, stores the documents in a compressed form, called signatures, which can be searched quickly using the parallel processors of the Connection Machine [Stanfill 86]. The signatures are stored in the RAM of the machine thereby assigning a few documents to each processor of the machine. Each word in the query is then broadcast to all the processors, and a score is kept for each document to reflect the number of words and phrases that matched. Weighting is done based on crude proximity and occurrence frequency. The resulting search results have been found to be useful to end-users.

As the dissemination of information becomes easier, questions of ownership, copyright, and theft of data must be addressed. These issues confront the entire information processing field, and are particularly acute here. The WAIS system is designed to keep control of the data in the hands of the servers. A server can choose to whom and when the data should be given. Documents are distributed with an explicit copyright disposition in their internal format. This is not to say that theft can not occur, but if a client starts to resell another's data, standard copyright laws can be invoked. By keeping the control of the distribution of works with the creators, many of the problems of copyright do not arise.

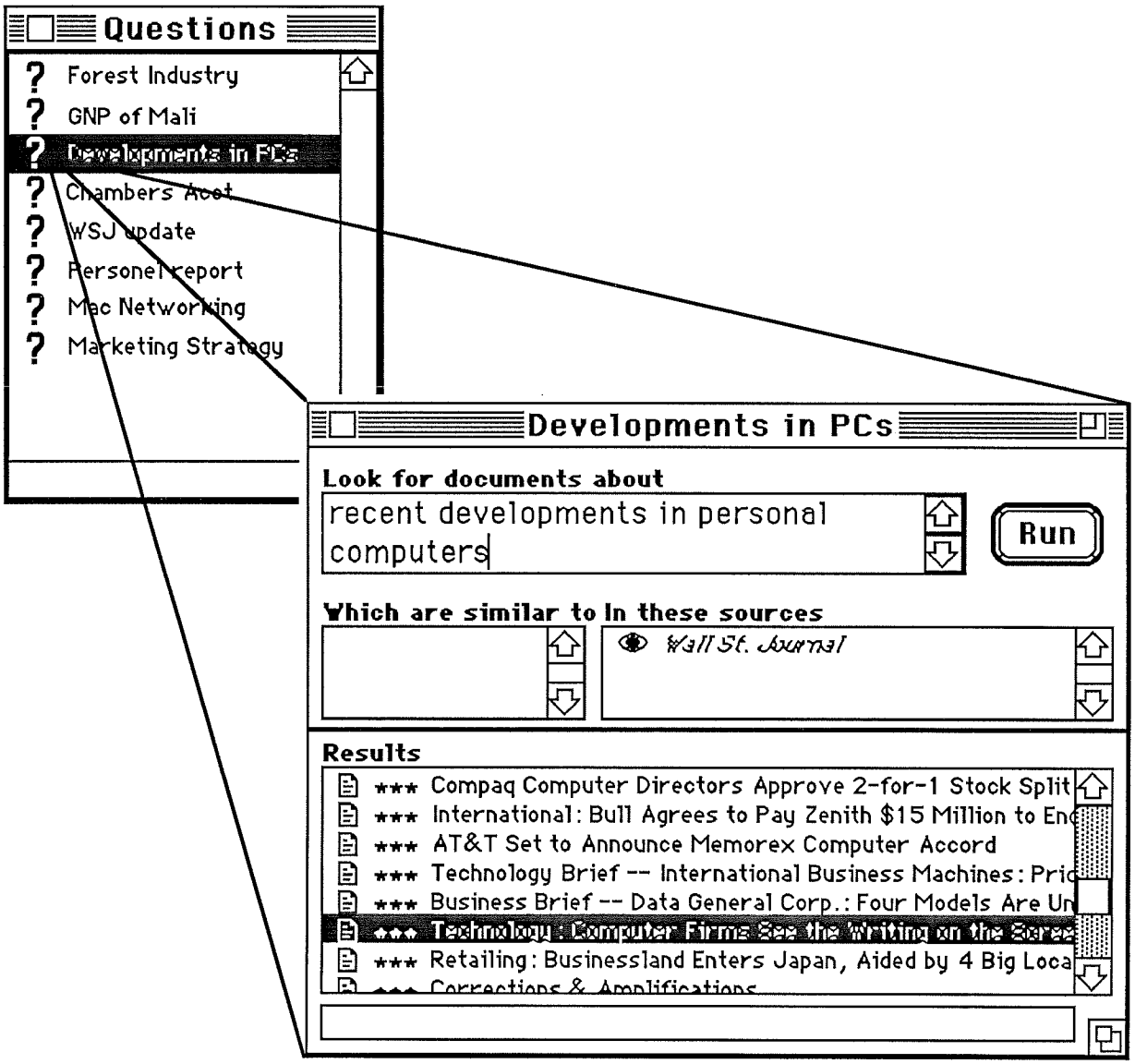

Figure 6: Opening a saved question which was automatically updated in the back ground, and contains new data.

2.5 Rerunning Questions - A Personal Newspaper

In addition to providing interactive access to information, the WAIS system can also be used as a rudimentary personal newspaper to alert its user when new documents are available on a subject that might be of interest. In the library literature, this is referred to as selective dissemination of information (SDI), and many manual, semi-automated, and automated systems have been implemented. Our initial implementation is to save interactive questions and automatically rerun them at periodic intervals checking if new documents were available. This technique has the advantage of hiding communication costs, using systems off hours, and finding potentially interesting information in a timely manor.

2.6 MultiMedia Database

The documents retrieved through WAIS may be any kind of file such as text, still graphics, motion pictures, or hypertext documents. The searching of the system is based on an initial quasi-natural language question and further relevance indications, but the server is free to use that information in any way to find appropriate documents. The protocol simply defines a document as a block of data and a type. The type is used by the client to determine how to display the document. A list of available types is part of the search response of each document. This allows clients to choose among a selection of types, and suppress documents whose types they can't display. Alternatively they can simply store the documents to their local disk for latter processing. Our initial X windows clients are able to use other programs to display graphic data such as TIFF and GIF. The Macintosh client can display PICT images and text, but can theoretically download any type of file.

Non-textual data is indexed in one of two ways. If the data includes an embedded description (e.g. TIFF), the description is used for indexing. Otherwise an external description is indexed. When a search identifies the description file as a suitable response, the multi-media data is returned instead.

2.7 The Directory of Servers

To find sources of information in a distributed environment, we used a "directory of servers" which is a database of documents describing other servers. In response to a query, the database is searched, returning a list of documents which match the query. Instead of text documents, however, it takes advantage of the mixed type capabilities of WAIS to return a structured document with many specific fields for cost and contact information. This capability will become more important as the number of servers increases.

For example, suppose you needed information concerning the current gross national product of Mali, but had no idea on which server to find it. You could first ask the directory of servers for "information about the current economic condition of Mali." The directory will take the words in the query and find descriptions of the servers that contain those words. It might then return several documents. The World Factbook for instance, might appear because of a match on "economic condition". This source description could then be used as the source field of another question. This time, the system would contact the World Factbook, ask for the information, and possibly return a document with a description of Mali [World Factbook 74].

Additionally, the directory of servers provides a means for information providers to advertise the availability of their data. When a new source becomes available, the developers can submit a textual description, along with the necessary

information for contacting the server. This information is added to the directory, and becomes available to the public by the searching interface.

3. The Prototype WAIS System

In the fall of 1990 we installed an experimental WAIS system at Peat Marwick. The prototype was used by 20 users in 6 cities. Peat Marwick utilized corporate data in Montvale NJ, and Dow Jones information in Princeton NJ. The system was run successfully for 6 months with good user reactions.

KPMG is an example of an information intensive company. Their role as consultants requires that they are aware of new products, market fluctuations, changing laws, internal regulations, and competition. In addition, as a large organization, there is considerable internal information such as company contacts, bids, reports, and resumes which can be leveraged. Furthermore, their distribution in 40 countries, with 200 offices in the US alone, makes them a prime candidate for wide area information technology.

The primary users were located in San Jose and connected by 56kbaud and 9.6kbaud circuits to the servers in New Jersey. The 20 managers and partners in the Peat Marwick's accounting division used an 8192 processor Connection Machine system for serving reports, proposals, resumes, contacts, accounting manuals, the Peat Marwick Audit Manual, Management Guide, and Professional Development Courses, documents from the Financial Accounting Standards Board, the Government Accounting Standards Board, and the American Institute of Certified Public Accountants, and a tax library. The data were separated into 12 different databases, which could be searched separately or in any combination. There was also a virtual database consisting of all of these sources.

The connection to Dow Jones provided access to 1 Gigabyte of data, running on a 32K processor Connection Machine. The data consisted of a year of the Wall Street Journal, Barons, and 400 magazines. Each of the approximately 250,000 articles was a separate document. The ability to search personal data was not available at the time of the experiment.

3.1 Lessons Learned

The search technology performed well for finding useful data for users given little instruction. The speed of the searches (usually between 2 and 10 seconds) depended on the communication speed since the search took much less than a second. When the response time was greater than 10 seconds, the users voiced complaints, but in general they were very pleased with the search results. The ability to execute searches without prior training, and without in-depth knowledge of the database was essential to the users. Relevance feedback was used frequently and effectively by users who were aware of it's existence. Not all users realized it was available however. This is an opportunity for user interfaces to improve. For example, relevance feedback could be performed automatically on any document which the user chooses to view. This would result in a kind of automatic, dynamically linked hypertext system, where every document is "linked" to all similar documents.

The Macintosh user interface (WAIStation) also performed well in ease of use and adaptability. With a single demonstration, most users were able to execute searches and save their results. Left only with the manual, new users took 15 to 30 minutes to feel comfortable with the system. The ability to transparently search local and remote databases was greatly appreciated as reported in user feedback forms. The biggest problem we had with the interface was in

implementing the TCP and modem connections from the Macintosh. The automatic updating feature of WAIStation was rarely used and needs more work to make it more obvious and give better feedback when documents are found.

Wide area communications proved to be a difficult part of the project due to our resistance to use leased lines based on future cost projections. The original plan called for linking San Jose and Montvale with Shiva Telebridges™ running at 9600 baud on a normal phone line. This approach did not prove reliable, nor give us reasonable performance. We ended up replacing this link with a dedicated 56kbaud line attached to a SyncRouter (Engage Communications™). The dedicated line was highly reliable and 56kbaud was fast enough to support many active users of the system, and maintain an interactive feel in both search and retrieval.

Organizing and formatting the data for display on the client workstation proved to require more effort than we expected. The current Macintosh client is only capable of displaying ascii text and PICT format picture files. This meant that the corporate data, which consisted primarily of word processor files, had to be converted to ascii. Since the conversion was not perfect, some documents required a small amount of manual reformatting. This is obviously unacceptable in a production system. A more attractive solution might be to build a client which can display the most common document formats, and which can call on other applications to display formats it doesn't understand. This approach will become easier as document filters (e.g. Claris™ XTND) and interprocess communication become more common. This will also make it possible to index and store the original document, rather than an ascii shadow.

As the searchable Peat Marwick corporate collection grew, the users wanted to search just parts of the database. The natural divisions for the users were the original sources of the text, such as training manuals or government legal texts.

In summary, we found the users were pleased with the system, and some used it many times each day. It appears that there is a market for end user search systems, and that the technology is ready. The weak link seems to be communication infrastructure.

4. Conclusion

In developing the WAIS system, the participating companies have demonstrated that current hardware technology can be effectively used to provide sophisticated information retrieval services to novice end-users. How this might effect information providers is not yet understood. The users at Peat Marwick found the technology useful for day-to-day tasks such as researching potential new accounts and finding resources within their own organization. Since these tasks are not restricted to the accounting and management consulting industries, we are optimistic that this type of technology can be fruitful and productive in many corporate settings.

The future of this system, and others like it, depends upon finding appropriate niches in the electronic publishing domain. Potential uses include making current online services more easily accessible to end-users, and allowing large corporations to access their own internal data more effectively. It is also possible that near-term development will focus on a single professional field such as patent law or medical research.

Acknowledgments

The design and development of the WAIS Project has been a collective effort, with contributions and ideas coming from many people. Among them:

Apple Computer: Charlie Bedard, David Casseras, Steve Cisler, Ruth Ridder, Eric Roth, John Thompson-Rohrlich, Kevin Tiene, Gitta Soloman, Oliver Steele, Janet Vratny-Watts.

Dow Jones News/Retrieval: Rod Wang, Roland Laird.

KPMG Peat Marwick: Chris Arbogast, Mark Malone, Tom McDonough.

Thinking Machines: Dan Aronson, Jonathan Goldman, Chris Madsen, Patrick Bray, Danny Hillis, Rob Jones, Barbara Lincoln, Gordon Linoff, Gary Rancourt, Sandy Raymond, Tracy Shen, Craig Stanfill, Steve Schwartz, Robert Thau, Ephraim Vishniac, David Waltz, Uri Wilensky.

References

[Davis 90] F. Davis, B. Kahle, H. Morris, J. Salem, T. Shen, R. Wang, J. Sui, and M. Grinbaum, "WAIS Interface Protocol Prototype Functional Specification," Thinking Machines, April 1990. Available via anonymous ftp: /pub/wais/doc/wais-concepts.txt@quake.think.com or wais server wais-docs.src.

[Egan 89] D. Egan, J. Remde, L. Gomez, T. Landauer, J. Eberhardt, and C. Lochbaum, "Formative Design-Evaluation of SuperBook," ACM Transactions on Office Information Systems, January 1989.

[Erickson 91] T. Erickson and G. Salomon, "Designing a Desktop Information System: Observations and Issues," Proceedings of the ACM Human Computer Interaction Conference, 1991. ACM Press, April 1991, pp. 49-54.

[Ginther-Webster 90] K. Ginther-Webster, "Project Mercury," AI Magazine, 1990, pp. 25-26.

[Kahle 89] B. Kahle, "Wide Area Information Servers Concepts," Thinking Machines technical report TMC-202, November 1989. Available via anonymous ftp: /pub/wais/doc/wais-concepts.txt@quake.think.com or wais server wais-docs.src.

[Kahle 91-1] B. Kahle and H. Morris, "Source Description Structures," February 1991. Available via anonymous ftp: /pub/wais/doc/source.txt@quake.think.com.

[Kahle 91-2] B. Kahle and H. Morris, "Document Identifiers or International Standard Book Numbers for the Electronic Age," May 1991. Available via anonymous ftp: /pub/wais/doc/doc-ids.txt@quake.think.com.

[Malone 86] T. Malone, K. Grant, and F. Turback, "The Information Lens: An Intelligent System for Information Sharing in Organizations; In Human Factors In Computing Systems," CHI'86 Conference Proceedings (Boston, MA); ACM: New York, 1986; pp 1-8.

[NISO 88] "Z39.50-1988: Information Retrieval Service Definition and Protocol Specification for Library Applications," National Information Standards Organization (Z39), P.O. Box 1056, Bethesda, MD 20817. (301) 975-2814. Available from Document Center, Belmont, CA. Telephone 415-591-7600.

[Salton 71] G. Salton, "The Smart Retrieval System - Experiments in Automatic Document Processing," Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1971.

[Salton 83] G. Salton and M. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, New York, 1983.

[Stanfill 86] C. Stanfill and B. Kahle, "Parallel Free-Text Search on the Connection Machine System," Communications of the ACM 29,12, December 1986, pp.1229-1239.

[Stanfill 91] C. Stanfill, "Massively Parallel Information Retrieval for Wide Area Information Servers," International Conference on Systems, Man, and Cybernetics, Charlottesville, Virginia, October 1991.

[World Factbook 74] Publishing Sciences Group, Acton, Massachusetts, 1974.